The Predictive Power of Subjective Probability Questions

de Bresser, Jochem; van Soest, Arthur

Publication date:
2017

Document Version
Early version, also known as pre-print

CentER

Discussion Paper

No. 2017-046

THE PREDICTIVE POWER OF SUBJECTIVE PROBABILITY QUESTIONS

By

Jochem de Bresser, Arthur van Soest

6 November 2017

ISSN 0924-7815
ISSN 2213-9532

TILBURG UNIVERSITY

The Predictive Power of Subjective Probability Questions*

Jochem de Bresser† Tilburg University‡ Netspar§

Arthur van Soest Tilburg University Netspar

November 6, 2017

Abstract

This paper evaluates the predictive validity of stated intentions for actual behaviour. In the context of the 2017 Dutch parliamentary election, we compare how well polls based on probabilistic and deterministic questions line up with subsequent votes. Our empirical strategy is built around a randomised experiment in a representative panel. Respondents were either simply asked which party they will vote for, or were asked to allocate probabilities of voting for each party. The results show that for the large majority of the respondents, probabilities predict individual behaviour better than deterministic statements. There is, however, substantial heterogeneity in the predictive power of the subjective probabilities. We find evidence that they work better for those with higher probability numeracy, even though probability numeracy was measured eight years earlier.

Key words: Subjective probabilities; predictive validity; probabilistic polling; elections

JEL-codes: D84, C81, C25

*This work is part of the research programme Innovational Research Incentives Scheme Veni with project number 451-15-018, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO). We thank Rob Alessie, Tabea Bucher-Koenen, Joachim Winter and the audience at the Research Workshop "Empirical Economics" at the Munich Center for the Economies of Aging for helpful discussion and comments.

†Corresponding author. Email: j.r.debresser@uvt.nl.

‡Tilburg University, P.O. Box 90153, 5000 LE Tilburg the Netherlands.

§Netspar, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

1 Introduction

In inter-temporal models of economic behaviour, decisions are driven by expectations as well as preferences. An individual's decision to save, for instance, is driven not only by patience or risk aversion but also by, e.g., the agent's subjective distribution of future income or even their survival probabilities. The fact that different combinations of preferences and beliefs can rationalize the same observed behaviour has spurred interest in the direct measurement of expectations in surveys (Manski, 2004). This literature has reached the conclusion that at least from a theoretical point of view, the best way to elicit beliefs is to ask respondents to report the probability that some future event will occur (see Manski, 2004). Such subjective probabilities allow for straightforward measurement of uncertainty and are more comparable across individuals than verbal qualitative statements (Manski, 2004). On the other hand, however, reported probabilities have been found to be affected by non-classical measurement error such as rounding (Manski and Molinari, 2010; Kleinjans and Van Soest, 2014). Many studies have demonstrated that subjective probabilities also have empirical validity: they correlate in plausible ways with background variables and help to predict future outcomes and decisions (see, e.g., the overview in Hurd, 2009). Until now there has been no direct evidence comparing the predictive power of subjective probabilities with that of the traditional way of eliciting intentions through deterministic questions.

This paper analyses data from an experiment in which respondents were randomly allocated to different types of questions that measure expectations regarding future decisions (so-called *choice expectations*, Manski, 2004, or intentions). The expectations concern the party an individual will vote for in the Dutch parliamentary elections of March 2017 and were elicited approximately three months prior to the election. We compare intentions elicited by deterministic items ("which party will you vote for?") with probabilistic intentions ("what is the probability that you will vote for party x?"). Since individuals do not face any restrictions on their actual voting behaviour, this is a clean case in which intentions and outcomes

can be compared without the need to model or make assumptions on exogenous events that may influence the actual outcome.

The idea of using subjective probabilities to elicit voting intentions goes back to Meier and Campbell (1979), Meier (1980) and Maas et al. (1990), but none of these studies compare probabilistic and deterministic approaches. Manski (2004) reports on a small pilot study for the 2000 U.S. presidential election and large scale probabilistic polls have been carried out for the presidential races of 2008, 2012 and 2016. Research on the latter two elections focused on the extent to which probabilistic polls anticipated the actual aggregate election outcome. Evidence has been mixed: it was one of the most accurate polls in 2012 (Gutsche et al., 2014), but substantially over-predicted the Republicans' share of the popular vote in 2016. The analysis of the 2008 elections reported by Delavande and Manski (2010) is closest to the present paper, because it considers the predictive power of verbal and probabilistic polling questions at the level of the individual. The authors show that combining both types of items improves the prediction of actual votes. However, doing so is costly, since it entails asking two sets of questions to elicit voting intentions. Delavande and Manski (2010) acknowledge that their research design, in which both probabilistic and verbal questions were posed in quick succession to all respondents, does not allow them to evaluate which type of question works best. After all, responses to the verbal questions may be affected by the probabilities that respondents reported previously. Our empirical strategy avoids this problem, since it is based on a large split-sample design.

Three features distinguish the present study from previous work. Firstly, we exploit a randomised experiment in a large, representative household panel that allows us to compare the predictive power of deterministic and probabilistic intentions in a clean way. In contrast to the research described above, panel members were exclusively assigned to either type of question. Secondly, while previous efforts focused on U.S. presidential races that effectively amount to binary choices, we analyse the more fragmented setting of parliamentary elections

in the Netherlands. On March 15 2017 the ballot listed 28 parties, 13 of which made it into the parliament. Such profusion of options presumably makes probabilities more powerful, since there is more scope for doubt experienced by undecided voters, particularly when the election is still some time off (three months, in our case). Finally, our data come from a long standing panel for which a lot of information has been collected in prior surveys. This allows us to relate the predictive power of reported probabilities to relevant background information, such as probability numeracy.

Our results, based on linear as well as multinomial discrete choice models, indicate that on average and for the large majority of the population probabilistic questions are substantially better predictors of actual votes than deterministic ones. Using linear models we find that an increase in the reported probability of voting for a party from 0 to 100% increases the likelihood of actually voting for that party by 46-79% in the deterministic sample, compared to an increase of 70-97% in the probabilistic sample. We show that this added power of probabilities can be attributed to the question format and not to systematic differences between samples. While there is little variation in the predictive power of deterministic intentions, estimates of a random coefficients discrete choice model point at substantial heterogeneity in the predictive power of subjective probabilities. They work very well for a large majority of the respondents (84%, according to our estimates), but perform worse than deterministic statements for a small minority (16% of the sample). This heterogeneity is related to probability numeracy: probabilities are better predictors for individuals with higher probability numeracy. This finding is in accordance with earlier studies demonstrating substantial heterogeneity in individuals' ability to work with probabilities and, in relation to that, the value of their answers to subjective probability questions for predicting actual behaviour (see, Armantier et al., 2015, and Binswanger and Salm, 2017).

In order to choose between a survey design with subjective probabilities or deterministic intentions, a trade off should be made between the benefits and the costs. We therefore also

briefly consider potential costs, due to the larger burden on the respondents. We find that the survey with 14 subjective probabilities takes significantly longer than the same survey with 14 deterministic intentions questions, with a difference at the median of slightly less than 2 minutes. We find no significant difference between the two designs in respondent evaluations of the difficulty or the attractiveness of the survey. We conclude that these costs are dominated by the much larger predictive power of probabilities at the level of the individual.

The paper is organized as follows. Section 2 describes the data and presents some descriptive statistics. Sections 3 and 4 contain our results regarding the predictive power of deterministic and probabilistic intentions. We first estimate linear and non-linear models that demonstrate that subjective probabilities provide additional power to predict each individual's actual voting behaviour. Using one of our models, we then construct an index for predictive power and relate it to probability numeracy. Section 5 briefly analyses the additional respondent burden of subjective probability questions. Section 6 concludes.

2 Data

2.1 The LISS panel

The LISS panel is a large household panel, consisting of approximately 8000 individuals in 5000 households that are broadly representative for the Dutch population (Van der Laan, 2009; De Vos, 2010).[1] Households are selected randomly by Statistics Netherlands from the complete registry of all Dutch non-institutionalized households. Surveys are administered online, and selected households receive a simple computer and an Internet connection if they do not have a computer or Internet access.

[1]More information on the panel, including code books for all available data and instructions on how to obtain access, can be found on www.lissdata.nl.

Longitudinal information on a wide range of socio-economic and demographic topics is collected on a yearly basis in so-called "core" surveys. In addition, researchers can design their own questionnaires on specific topics. Our analysis combines data from four different surveys. The votes cast by respondents in the parliamentary elections of March 15 2017, our main outcome variable, were collected in an "exit poll" during the two weeks immediately following the election (between March 16th and 30th). Voting intentions for the same election were collected in the core *politics and values* survey of December 2016, approximately three months before the election.[2] Background variables are obtained from the *household box* of that month. Finally, in part of our analysis we use a "probability numeracy" variable calculated from items included in a one-off *disease prevention* survey that was designed by Katie Carman and Peter Kooreman and fielded in September of 2008 (see Bruine de Bruin and Carman, 2012, and Carman and Kooreman, 2014). Unfortunately, a more recent numeracy measure for the LISS respondents is not available.

2.2 Probabilistic poll

Crucially for this study, voting intentions were measured differently in 2016 compared to previous years. Inspired by the probabilistic polls for the 2012 and 2016 U.S. presidential elections,[3] an experiment was set up to compare responses to different types of polling questions. All respondents were asked to report their voting intentions in two steps. First, they were all asked in the same way to indicate the probability that they would vote:

If parliamentary elections were held today, what is the percent chance that you will vote?

Please fill in a percentage between 0 and 100:

0..100

[2]Respondents were invited to take the survey early December. Those who did not take the survey in December got another invitation in January 2017; only a small minority used this opportunity.

[3]See Gutsche et al. (2014), www.alpdata.rand.org/?page=election2012 and www.cesrusc.org/election/

Second, respondents forecasted which party they would vote for conditional on voting. One random half of the panel received a single deterministic question:

If parliamentary elections were held today, for which party would you vote?

[If $\Pr(vote) = 0$*: I would not vote], VVD (liberal party), ..., VNL, Another party, Blank*

This is the usual way voting intentions are measured in LISS. The answer options are the 14 parties represented in parliament at the time of the survey, any other party, and not casting a vote on any of the parties ("Blank"). Respondents who gave a 0% probability of voting at all in the previous question, got an additional option "I would not vote".

The other half of the sample were asked to assign probabilities to voting for different parties, voting "Blank", or not voting at all:

If parliamentary elections were held today, what is the percent chance that you will vote for each of the following parties? Total probability should add up to 100%.

[if $\Pr(vote) = 0$*: I would not vote], VVD (liberal party), ..., VNL, Another party, Blank*

In order to help respondents answer these questions in a logically consistent way, all parties were shown on a single screen and the total probability mass that they had already distributed was shown at the bottom. Respondents did not have to assign 15 (or 16) probabilities explicitly: fields left empty were counted as zeros. Moreover, respondents could not proceed to the next question in the survey if they provided probabilities outside the 0-100 interval or if their probabilities did not add up to 100%.

The "treatment" deterministic or probability questions was assigned completely randomly. As a consequence, the two treatment groups are similar in terms of observable characteristics (see the balance tests reported in Appendix A).

For our analysis, we compute the unconditional probabilities to vote for each party, combining the (unconditional) probability of voting at all with the conditional probabilities of voting for each of the parties given voting:

$$\Pr(\text{vote party x}) = \Pr(\text{party x}|\text{vote}) \times \Pr(\text{vote}) \tag{1}$$

Moreover, "no vote" is added as the remaining possible outcome (with probability $1 - \Pr(\text{vote})$). These unconditional probabilities will be analysed in relation to the actual voting behaviour observed in the exit poll. As explained above, half of the respondents got deterministic questions, restricting their probabilities $\Pr(\text{party x}|\text{vote})$ to zero or one. On the other hand, all respondents report their subjective probability to vote, $\Pr(\text{vote})$, as a probability. This implies that the unconditional probabilities can also take on values between 0% and 100%. We use the fact that there is no difference in elicitation method for "no vote" as a placebo treatment, since there is no reason to expect any difference in the predictive power of intentions across the two treatment groups for the "no vote" outcome.

2.3 Descriptive statistics

Actual vote

The Dutch political landscape around the time of the 2017 parliamentary elections was highly divided and voters could choose among 28 parties on the ballot. In addition, voters could show up at a voting bureau but not cast a (valid) vote on any of the parties (the "blank" option). At the time of our first survey (December 2016), the definitive list of parties on the ballot was not yet known but the parties not yet represented in parliament were not expected to attract many votes. In the survey we therefore only listed the 13 parties already represented in parliament at that time (including two new parties started by members of parliament who left their party during the term), as well as an option "other party". In the

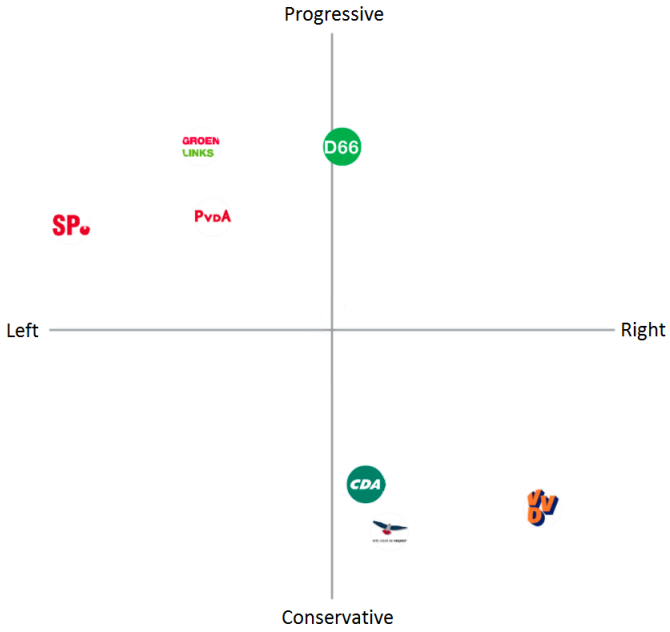

Figure 1: The Dutch political landscape in March 2017 (source: presentation by dr. André Krouwel, available here)

analysis, we combine the six smallest parties among these 13 with the original "other party" and the "blank" option into a larger "other" category to generate a multinomial outcome with nine options: "no vote" (or, to be more precise, no show up), a vote on one of the seven largest political parties, and "other" (a vote on another party or a "blank" vote).

Figure 1 shows where the seven major parties are located in ideological space, following the common two-dimensional party characterization of, e.g., Marks et al. (2006) and Van Kersbergen and Krouwel (2008). The horizontal axis labeled *left/right* reflects the economic dimension, expressing the distinction between egalitarian parties that favor extensive redistribution and regulation (*left*) versus parties with a more laissez-faire ideology (*right*). The vertical axis shows a non-economic dimension, with parties that favour cultural liberalism and openness at the top (*progressive*, often labeled GAL (green, alternative and libertarian)), and parties that favour restrictive immigration policies (*conservative* or TAN (traditional, authoritarian, nationalist)) at the bottom. Hence, the economically liberal yet culturally

Table 1: Descriptive statistics of the outcome variable: actual vote in 2017 elections

a. Missing outcome (non-participation in exit poll)

	Overall					
	Mean	SD	Deterministic	Probabilistic	Difference	(SE)
Vote missing	0.09	0.28	0.09	0.08	-0.01	(0.008)
N	4349		2131	2218	4349	

b. Dependent variable: actual vote in 2017 elections (0% or 100%)

		Overall sample					
	Population[a]	Mean	SD	Deterministic	Probabilistic	Difference	(SE)
VVD (liberal)	17.4	18.3	38.7	18.5	18.1	-0.4	(1.22)
Other party	14.1	13.9	34.6	13.8	14.1	0.3	(1.08)
CDA (christian)	10.1	13.4	34.1	13.7	13.2	-0.5	(1.09)
D66 (prog. lib.)	10.0	12.2	32.7	11.3	13.1	1.8*	(1.03)
GL (green)	7.5	9.7	29.5	9.3	9.9	0.6	(0.94)
PVV (populist)	10.7	9.2	28.9	9.6	8.8	-0.8	(0.90)
SP (socialist)	7.4	9.1	28.8	9.2	9.0	-0.3	(0.89)
PvdA (labour)	4.7	7.1	25.7	6.8	7.4	0.6	(0.81)
No vote	18.1	7.1	25.7	7.7	6.5	-1.3	(0.82)
N		3978		1936	2042	3978	

Chi-squared test for equality of vote distribution across treatments: $\chi^2\,(8) = 7.02$, p-value $= 0.54$.
Standard errors in parentheses; clustered at household level (3275 clusters for missing DV model, 3027 clusters for DV). $^*p < 0.1$

Notes

[a] Percent of the population that was eligible to vote (NOT a percentage of the vote).

conservative VVD can be found in the bottom right corner and the progressive leftists of GroenLinks (Green Left, GL) in the top left. The Partij Voor de Vrijheid (Freedom Party, PVV) led by Geert Wilders was the most conservative party in the GAL/TAN dimension, yet its economic ideas are middle-of-the-road.

Table 1 reports descriptive statistics of actual voting behaviour, as reported in the exit poll in the two weeks after the elections. Panel a. shows that only 9% of panel members who were eligible to vote and participated in the voting intentions survey in December 2016 did not participate in the exit poll. This fraction is almost identical for both treatment groups. Panel b. compares voting behaviour reported in our "exit poll" survey with voting behaviour of the complete population. The liberal VVD received the largest share of the vote, 18.3% in the sample and 17.4% in the population. The category "other party" got approximately 14% of the votes, both in the sample and the population. While this large number may suggest

splitting up the category "other" into its constituents, doing so would create outcomes that are rarely chosen since this category is comprised of many rather small parties (and the "blank" option).

The Christian democrat CDA is the second largest party in the sample at 13%, followed by the progressively liberal D66 at just over 12% (with corresponding population figures around 10% for both). The greens (GL) and the socialists (SP) received between 9 and 10% of the votes in the panel, as did the populist PVV. While the greens and socialists did better in the panel than in the population, the opposite is true for the PVV, which became the second largest party with the support of 10.7% of the vote-eligible population. The smallest individual party in our analysis is the labour party (PvdA), with 7% of the votes in the sample and less than 5% in the population. With the exception of the PVV, the ranking of parties is the same in the sample as in the population. On the other hand, there is a large and salient difference between sample and population when it comes to the proportion that did not vote at all: 18% in the population and only 7% in the sample. To put this discrepancy of 11 percentage points (pp) in perspective, Delavande and Manski (2010) found that for the presidential elections in 2008, turnout in the American Life Panel (ALP) was 28pp higher than in the complete population. We cannot say whether the high reported turnout in the LISS is due to selection, an effect of panel participation on the likelihood of voting, or simply misreporting. For our analysis this is not really relevant, since we compare two randomised treatment groups within the LISS sample. A Chi-square test does not reject the null hypothesis that the voting patterns (columns *deterministic* and *probabilistic* in the table) are the same for the two groups with different treatments (p-value = 0.54).

Intentions

Figure 2 shows the distribution of reported probabilities for the "no vote"-option for those respondents reporting a positive probability. Its two panels correspond to the sub-samples

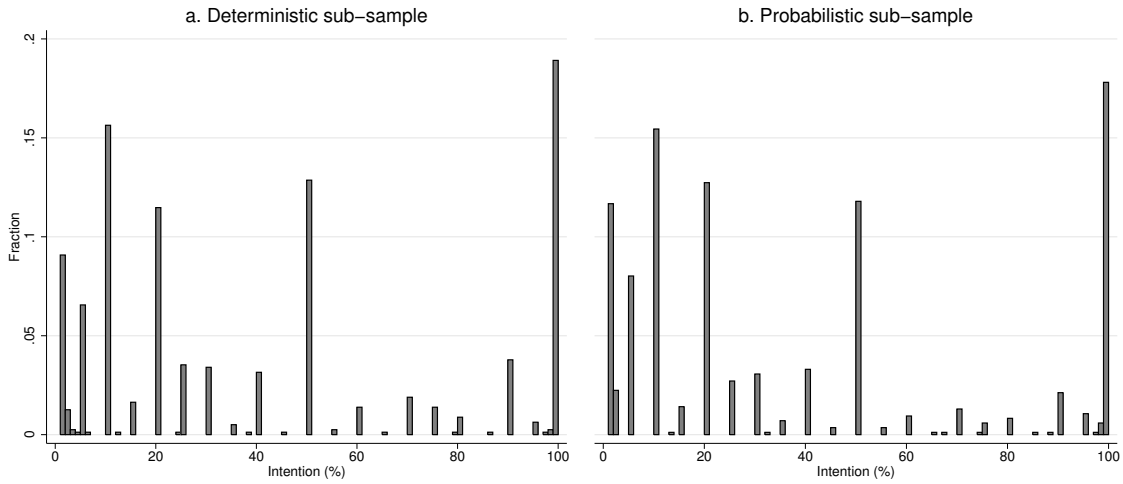

Figure 2: Histograms of probabilities for "no vote" excluding zeros

that received probabilistic and deterministic questions. Note that for "no vote" both samples received the same, probabilistic, question, so we can use this probability for a placebo test on systematic differences in accuracy of predictions across the two sub-samples that are not due to different ways of eliciting expectations. Indeed, the two histograms look similar and we cannot reject the null hypothesis that the two distributions of "no vote" intentions are the same: a Kolmogorov-Smirnov test for equality of the distributions yields a p-value of 0.85.

The histograms also show bunching of reported probabilities at multiples of 10 that is common for data on subjective probabilities (cf., e.g., Manski and Molinari, 2010; Kleinjans and Van Soest, 2014). There is some additional heaping at 50 percent, a value that has been associated with focal answers not reflecting genuine uncertainty (Bruine de Bruin et al., 2000). This suggests that the observed degree of bunching can partly be explained by rounding, but some of it may also be due to "epistemic" uncertainty – the inability to translate uncertainty into meaningful probabilities. In our study we do not aim at isolating rounding errors or focal answers. We take the reported probabilities at face value and analyse how well they predict actual behaviour.

Table 2 contains descriptive statistics of voting intentions. It presents the overall means,

Table 2: Descriptive statistics of intentions: stated intention to vote (three months before the election; 0-100%)

	Deterministic				Probabilistic			
		Fraction equal to				Fraction equal to		
	Mean	0	1-99	100	Mean	0	1-99	100
VVD (liberal)	13.7	0.85	0.04	0.10	13.7	0.67	0.29	0.04
Other party	18.1	0.78	0.09	0.12	14.6	0.60	0.36	0.04
CDA (christian)	7.9	0.92	0.02	0.06	9.4	0.72	0.25	0.03
D66 (prog. lib.)	9.1	0.90	0.03	0.07	11.1	0.65	0.34	0.02
GL (green)	7.1	0.92	0.03	0.05	8.0	0.73	0.26	0.02
PVV (populist)	13.9	0.84	0.06	0.10	11.5	0.75	0.20	0.05
SP (socialist)	6.5	0.93	0.03	0.04	7.4	0.75	0.23	0.02
PvdA (labour)	6.4	0.93	0.03	0.04	8.3	0.72	0.26	0.02
No vote	17.2	0.59	0.33	0.07	16.1	0.58	0.34	0.07
N	1936				2042			

as well as means and other summary statistics for the sub-samples that received probabilistic and deterministic questions. In the sub-sample that faced a deterministic choice between parties, only 2-9% of the probabilities are not equal to either 0 or 100 percent. This is because all respondents in this treatment who report a 0 or 100 percent probability of not voting at all, automatically get probability 100 or 0 for each party. In contrast, the probabilistic sub-sample exhibits substantial variation across political parties in the fraction of intermediate probabilities. Only 20% doubt between voting for the populist PVV and some other option, while 34% considers voting for the progressive liberals of D66 but is not certain yet.

Histograms of reported intentions to vote for the three most prevalent options are presented in Figure 3. We limit the sample to respondents who received probabilistic questions and reported a strictly positive probability. Bunching at multiples of 10 is evident for all options. There is also some bunching at other values (e.g., 18, 45) that is due to the fact that the probabilities in this figure are computed as the product of the probability to vote and the conditional probability to vote for a given party, which both have their own bunching due to rounding. Nonetheless, rounding remains visible, as well as bunching at 50/50 that

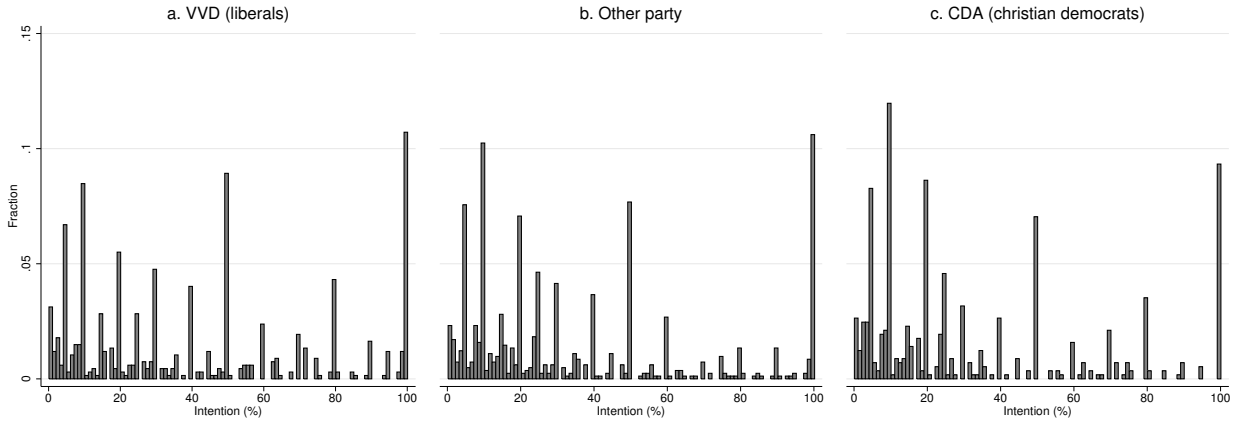

Figure 3: Histograms of probabilities for three most prevalent options excluding zeros (probabilistic sub-sample)

could be due to epistemic uncertainty, as in Figure 2.

For respondents in the probabilistic subsample, we can define the "consideration set" as the set of parties to which a respondent assigns a positive probability. Figure 4 summarizes the size of the consideration sets of respondents in the probabilistic sample for those individuals who report a positive probability that they will vote (93% of the sample).[4] At approximately three months before the election, only 30% of respondents have already fully made up their mind (100% for one specific option). About 47% doubt between two or three options and 11% spread their probability mass over four. All in all, panel members use the flexibility of the probabilistic questions to express a plausible level of uncertainty.

In order to facilitate comparing intentions with actual votes in the aggregate, Figure 5 combines the overall actual and intended vote shares for all alternatives. Actual votes are on the vertical and intentions on the horizontal axis. If predicted and actual vote shares were exactly equal, all circles would lie on the 45 degree line. This is apparently not the case. The probabilistic expectations (solid circles) are closer to the diagonal than the deterministic expectations (hollow circles) in 7 out of 9 cases. Moreover, the differences are significant at the

[4]The "no vote" option is not counted here, and non-zero probabilities for more than one party in the "other" category are counted as one.

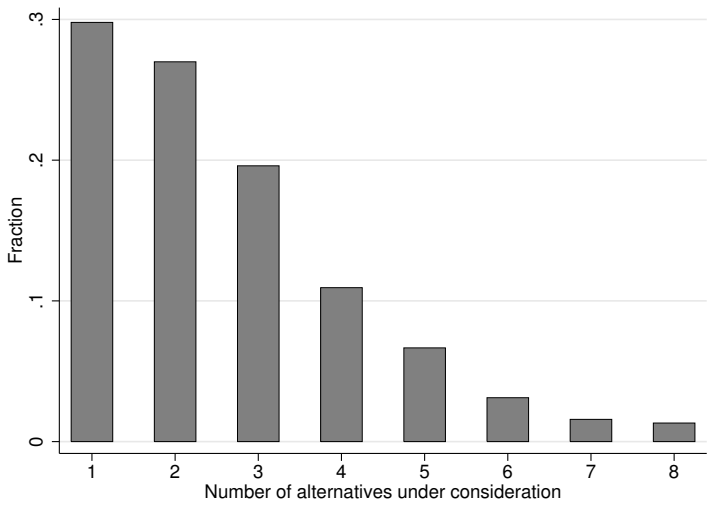

Figure 4: Distribution of size of consideration set: the number of parties to which respondent assigns positive probability (probabilistic sub-sample)

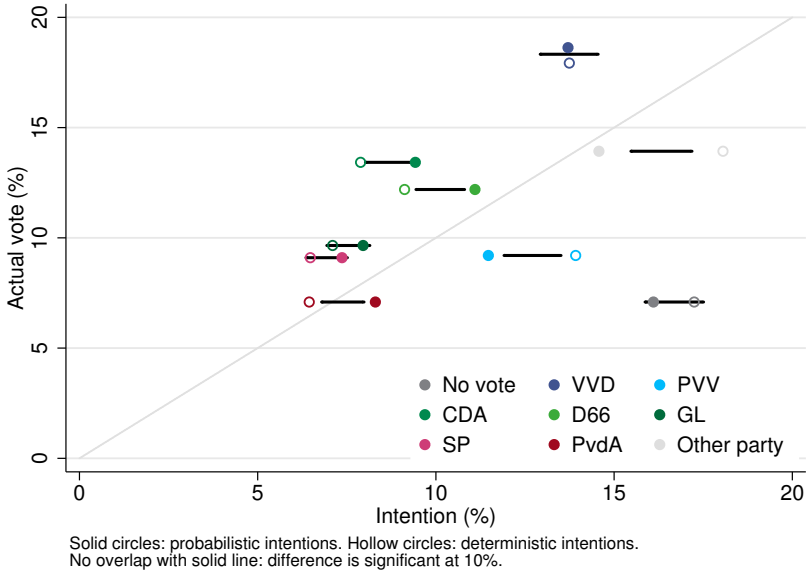

Figure 5: Aggregate intentions and actual votes

10% level for 4 of the 7 options for which probabilities outperform deterministic statements. (The remaining cases are VVD for which probabilistic and deterministic aggregates are almost the same, and PvdA where both are on different sides of the 45 degree line, with the deterministic aggregate somewhat closer to it than the probabilistic forecast.) This suggests that the probabilities give better predictors of aggregate behaviour than the deterministic answers.

The focus of our paper, however, is not the aggregate level but the predictive power of subjective probabilities at the individual level. Figure 6 provides a first impression of the relationships between intentions and actual votes. The graphs present kernel regressions of an indicator for choosing each alternative on the reported probabilities. In the first "no vote" graph, there are separate regressions for the two treatments, but this is hard to see since the lines and confidence bands overlap. This is reassuring, since the intention not to vote was elicited by the same question in both samples. The fitted lines lie substantially below the diagonal: respondents underestimate the probability they will vote at all levels of stated expectations. The fraction of people who abstain from voting is below 20% all the way up to a reported intention around 0.8. The rate of non-voters rises among those who report being nearly certain they will not vote, but never beyond 50%. This is not due to smoothing: among those who reported complete certainty that they will not vote, 47% do participate in the ballot. We can interpret this as a macro-effect: events and developments between December 2016 and March 2017 (or even the weather on the day of the elections) have stimulated interest in the elections and have increased voting rates across the board.

For the other outcomes, similar kernel regressions are performed for the probabilistic sample only. We visualize the relationship differently for the deterministic treatment, because of the scarcity of probabilities unequal to either zero or one. For this subsample, the plots show the mean probabilities of choosing each alternative given that the reported intention

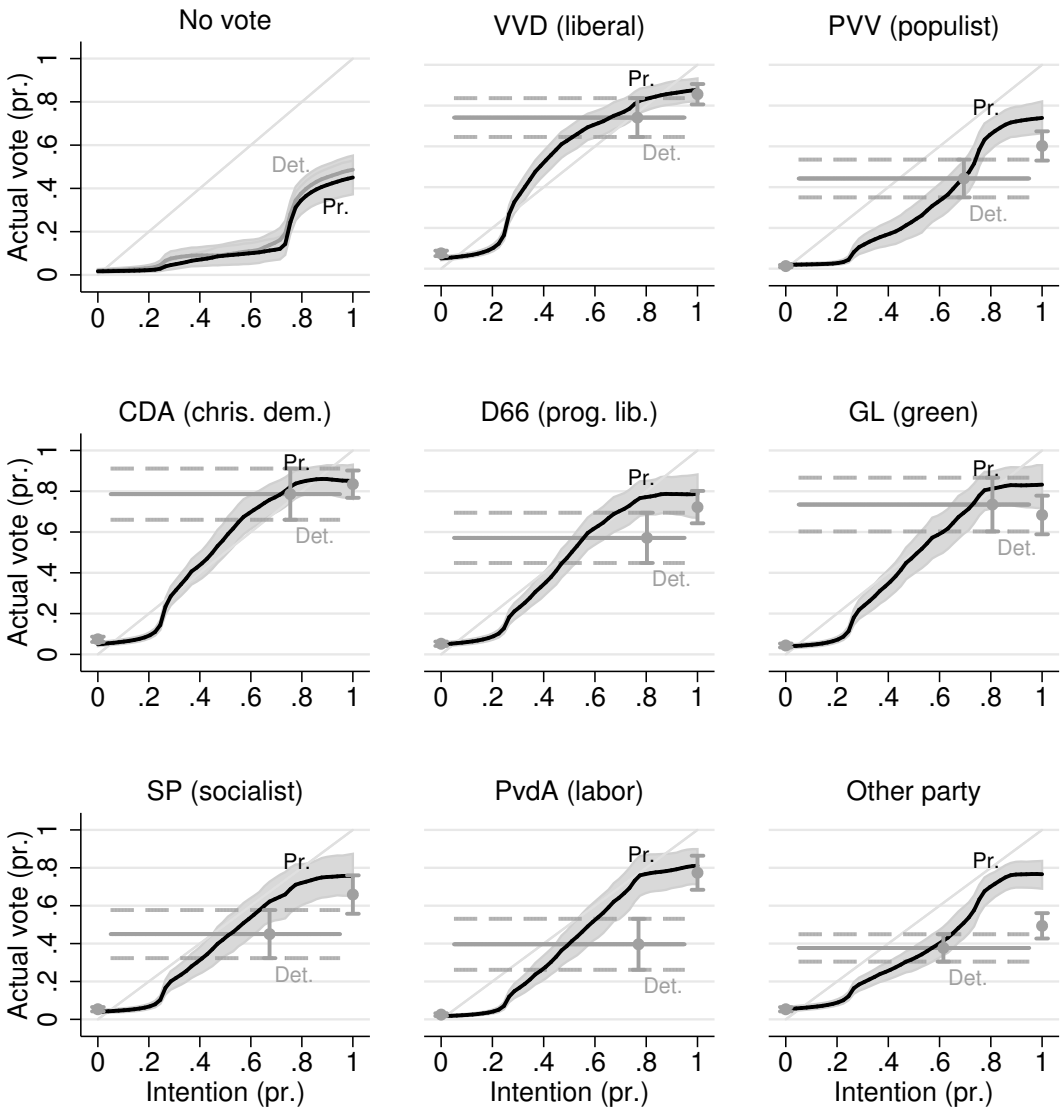

Figure 6: Kernel regressions of actual vote on voting intention (bandwidth 12pp)

falls in one of the three categories included in Table 2: zero, one, or some probability in-between (which we anchor on the average probability between 0.01 and 0.99 in the figure).

In both subsamples, intentions appear to be better predictors of voting for a specific party than of the decision not to vote. The probabilistic questions show a clear S-shaped pattern that is similar for all parties: the likelihood of voting for a party is close to zero for intentions below 0.2, then increases almost linearly to around 80% for intentions around 0.8, and then levels off. Intermediate levels of intentions understate the likelihood of voting for the liberal VVD, the largest party in our sample (60% of those who report a 50/50 chance of voting VVD end up doing so). The opposite applies to the populist PVV and the "other" option (only 30% of those who predict a 50/50 chance chose those options). The voting behaviour of the deterministic and probabilistic sub-samples is similar for low levels of intentions: if someone in the probabilistic sample assigned a very low probability to a given party, the chances that he or she voted for that party are very low. Similarly, if someone in the deterministic sample intended not to vote for a specific party, the chances he or she voted for that party are equally low. On the other hand, there is a notable difference between the two samples at the other end of the intentions range. Among those in the probabilistic sample with probability close to one to vote for a specific party, the fraction who indeed voted for that party is typically larger than the same fraction among those the deterministic group who intended to vote for that same party. This is especially true for the PVV and the "other" parties, for which only 60% and 50% voted for the party they indicated, respectively.

3 Regression models

3.1 Linear models

We first present estimates of separate linear probability models for each of the nine actual voting outcomes; see Table 3. Panel a. contains our baseline estimates, using all the data

Table 3: Linear probability models of actual vote (0% or 100%)

a. Baseline (no controls included in the models)

		Parties							
	no vote	VVD	PVV	CDA	D66	GL	SP	PvdA	other
Intention (%)	0.428***	0.791***	0.572***	0.782***	0.658***	0.680***	0.586***	0.686***	0.456***
	(0.0319)	(0.0243)	(0.0326)	(0.0323)	(0.0369)	(0.0412)	(0.0452)	(0.0411)	(0.0299)
Intention × prob. questions	-0.0349	0.180***	0.127***	0.176***	0.221***	0.218***	0.216***	0.163***	0.266***
	(0.0446)	(0.0387)	(0.0470)	(0.0483)	(0.0567)	(0.0609)	(0.0667)	(0.0593)	(0.0434)
Prob. questions	-0.232	-2.863***	-0.852*	-3.370***	-1.937**	-1.724**	-2.399***	-1.995***	-2.021**
	(0.539)	(0.871)	(0.464)	(0.813)	(0.782)	(0.681)	(0.713)	(0.503)	(0.796)
Constant	0.365	7.673***	1.641***	7.524***	5.255***	4.516***	5.453***	2.346***	5.552***
	(0.403)	(0.684)	(0.337)	(0.655)	(0.543)	(0.495)	(0.585)	(0.370)	(0.592)
N	3978	3978	3978	3978	3978	3978	3978	3978	3978
R-squared overall	0.24	0.47	0.43	0.38	0.34	0.35	0.26	0.43	0.27
R-squared deterministic	0.25	0.47	0.42	0.36	0.34	0.35	0.23	0.42	0.24
R-squared probabilistic	0.23	0.47	0.44	0.40	0.34	0.34	0.30	0.44	0.31

b. Probabilistic intentions replaced by mode (no controls included in the models)

		Parties							
	no vote	VVD	PVV	CDA	D66	GL	SP	PvdA	other
Intention (%)	0.428***	0.791***	0.572***	0.782***	0.658***	0.680***	0.586***	0.686***	0.456***
	(0.0319)	(0.0243)	(0.0326)	(0.0323)	(0.0369)	(0.0412)	(0.0452)	(0.0411)	(0.0299)
Intention × prob. questions	-0.0349	-0.0145	0.0411	-0.000952	-0.00733	0.0475	0.0448	-0.00548	0.151***
	(0.0446)	(0.0353)	(0.0453)	(0.0443)	(0.0500)	(0.0545)	(0.0622)	(0.0556)	(0.0420)
Prob. questions	-0.232	-1.019	-0.121	-1.866**	0.546	-0.271	-1.044	-0.564	0.524
	(0.539)	(0.901)	(0.451)	(0.828)	(0.781)	(0.685)	(0.714)	(0.498)	(0.799)
Constant	0.365	7.673***	1.641***	7.524***	5.255***	4.516***	5.453***	2.346***	5.552***
	(0.403)	(0.684)	(0.337)	(0.655)	(0.543)	(0.495)	(0.585)	(0.370)	(0.592)
N	3978	3978	3978	3978	3978	3978	3978	3978	3978
R-squared overall	0.24	0.45	0.43	0.38	0.33	0.35	0.25	0.43	0.27
R-squared deterministic	0.25	0.47	0.42	0.36	0.34	0.35	0.23	0.42	0.24
R-squared probabilistic	0.23	0.44	0.43	0.40	0.32	0.36	0.28	0.44	0.29

Systems of linear probability models are estimated as SUR models (allowing for dependence between error terms of different equations).
Robust standard errors in parentheses, clustered at household level (3027 clusters).
***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

available for the two sub-samples. The leftmost column corresponds to the "no vote" outcome. In line with the kernel regressions in Figure 6, the predicted probability of not voting increases from close to 0% for a reported intention of zero percent to 43% for an intention equal to 100%. The fact that the intention not to vote does not predict that outcome well is also apparent in the low R-squared of 0.24. Importantly, the intercept and the slope of this relationship do not differ significantly between the probabilistic and the deterministic samples (neither jointly, nor individually). This is in line with what we would expect since no-voting intentions are asked in the same way in the two sub-samples. Hence the placebo test is passed.

For the political parties, intentions are more predictive of behaviour than for the "no vote" option and there is added value for probabilistic questions relative to deterministic ones. For example, in the deterministic sample the predicted probability of voting for VVD increases from 8% for someone announcing not to vote for VVD to 87% for someone announcing he or she will vote VVD. In the probabilistic sample the intercept is almost 3pp lower and the slope 18pp steeper (the coefficient on the interaction *Intention* $\times$ *prob. questions*), so the corresponding increase is from 5% for someone with intention 0% to vote VVD to approximately 100% for someone with intention 100% to vote VVD. Furthermore, the R-squared is 0.47, approximately twice that for "no vote". While the predictability of the vote varies across parties, the general pattern described for VVD is found for all parties. In the deterministic sample the predicted probability of voting increases substantially if someone intends to vote for that party, from between 1.6 and 7.7% to between 51 and 87%. For the probabilistic sample, the intercept is always significantly smaller and the slope is always significantly steeper, and the predictions rise from between 0.35 and 4.9% for someone with intention 0% to vote for that party to between 71% and 100% at intention 100%. Deterministic intentions do worst for the "other party" outcome, but probabilistic intentions work much better here. Probabilistic intentions do worst (and have the lowest

added value) for the populist PVV. Still, even for this party, the subjective probabilities have significantly larger predictive power than the deterministic statements.

Panel b. of Table 3 presents estimates for the same regression models but with the probabilistic intentions transformed to match intentions elicited in the deterministic treatment. Since both samples received the same question for "no vote" we did not adjust those intentions and the estimates are identical to those reported in panel a. For the parties, we replace the conditional probabilities for the probabilistic sample by 100% for a unique mode, splitting probability mass evenly in case of multiple modes (which occurs for 15% of the observations), and 0% for the other options. The results show that this discretisation of probabilistic intentions largely removes their added value for prediction relative to the deterministic questions. The differences in slopes (i.e., the coefficients on the interactions) are reduced to close to zero and insignificant for all parties except "other" (for which the difference is almost halved to 15pp but remains significant). This demonstrates that the additional predictive power of subjective probabilities is almost completely due to the more detailed information that these probabilities provide. If this information is largely removed (largely because in the multiple modes case, the transformed probabilities are still more informative than the deterministic intentions), the additional predictive power is lost almost completely.

The models presented in Table 3 do not contain any other covariates, and thus rely on the randomisation to ensure that probabilistic and deterministic samples are comparable. Appendix C contains estimates of similar linear models that control for a wide range of covariates. The estimates of the slopes of probabilistic and deterministic intentions are virtually identical to those in Table 3. Moreover, the R-squared increases only slightly (never by more than 3pp), indicating that covariates such as gender, age and education have little explanatory power once we control for intentions.

Summarizing, we find clear evidence that subjective probabilities are much better in predicting individual behaviour than deterministic intentions. This added value is a con-

sequence of the finer response scale which provides additional information, and disappears when probabilities are transformed into modes. In the next section we turn to multinomial discrete choice models and investigate how the predictive power of subjective probabilities varies across individuals.

3.2 Multinomial choice models

Multinomial choice models account for the binary and joint nature of the nine actual voting outcomes. Table 4 contains estimates of two models. The first (panel a.) is a standard multinomial logit model with fixed coefficients. It contains alternative-specific constants and their interactions with the dummy for the probabilistic sample. Moreover, interactions with deterministic and probabilistic intentions are added, allowing the predictive power of deterministic intentions and the added value of probabilistic intentions to vary across parties (as in the linear models reported in Table 3). The estimates tell a similar story as the linear models. Firstly, the intention not to vote significantly predicts not voting and carries the same predictive power in both subsamples, as expected (the placebo test). Second and more importantly, for all political parties except one (PVV), intentions collected by means of probabilistic questions have significantly larger predictive power than the deterministic ones: the coefficients on the interactions between reported intentions and the probabilistic treatment dummy are always positive, and significant in all cases except PVV.

One way to increase the flexibility of the multinomial logit model and allow for heterogeneous treatment effects is to model key parameters as random coefficients. In order to keep the number of random coefficients manageable, we assume that in both subsamples, the effects of intentions are the same for all parties (but not for "no vote", for which the treatment is a placebo treatment). We estimate a random coefficients version of the multinomial logit model (often called mixed logit model) with two random coefficients: the coefficient on intentions and the coefficient on the interaction of intentions with the subjective probabilities

Table 4: Multinomial choice models of actual vote

a. Fixed coefficients

		Alternatives							
	Intentions	VVD	PVV	CDA	D66	GL	SP	PvdA	other
Intention (%)	0.0297***								
	(0.00251)								
Intention × prob. questions	-0.00264								
	(0.00366)								
Intention × party		0.00456	0.00540	0.00356	-0.00100	0.00146	-0.00483	0.00798**	-0.0118***
		(0.00330)	(0.00333)	(0.00373)	(0.00336)	(0.00355)	(0.00347)	(0.00355)	(0.00309)
Intention × prob. × party		0.0186***	0.00833*	0.0196***	0.0184***	0.0203***	0.0177***	0.0198***	0.0155***
		(0.00551)	(0.00498)	(0.00636)	(0.00578)	(0.00643)	(0.00592)	(0.00655)	(0.00482)
Prob. questions		-0.192	0.181	-0.273	0.0383	-0.0164	-0.178	-0.0969	-0.0268
		(0.295)	(0.345)	(0.302)	(0.307)	(0.313)	(0.306)	(0.342)	(0.304)
Constant		1.286***	-0.0441	1.308***	0.982***	0.830***	1.046***	0.197	1.182***
		(0.207)	(0.254)	(0.207)	(0.214)	(0.217)	(0.216)	(0.240)	(0.214)
Observations					35,802				
Individuals					3978				
Log-likelihood					-5273.95				

b. Mixed logit – independent normal mixing distributions

		Alternatives							
	Intentions	No vote	PVV	CDA	D66	GL	SP	PvdA	other
Means of parameters									
Intention (%)	0.0303***								
	(0.00145)								
Intention × prob. questions	0.0266***								
	(0.00289)								
Intention × no vote		-0.00136							
		(0.00340)							
Intention × prob. quest. × no vote		-0.0345***							
		(0.00605)							
Prob. quest. × no vote		0.247							
		(0.271)							
Constant		-1.354***	-1.235***	-0.0540	-0.334***	-0.448***	-0.449***	-0.949***	-0.651***
		(0.190)	(0.111)	(0.0761)	(0.0791)	(0.0840)	(0.0845)	(0.0997)	(0.0832)
Standard deviations of parameters									
Intention (%)	0.00856								
	(0.00551)								
Intention × prob. questions	0.0266***								
	(0.00384)								
Observations					35,802				
Individuals					3978				
Log-likelihood					-5309.04				

The dependent variable distinguishes between: no vote; VVD; PVV; CDA; D66; GL; SP; PvdA; other.
Robust standard errors in parentheses, clustered at household level (3027 clusters).
***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

treatment. In this way, we allow the predictive power of intentions in both subsamples to vary across respondents in a parsimonious way. (We also experimented with models with more random coefficients, but did not find substantial heterogeneity in other coefficients; estimates are available on request.)

Panel b. of Table 4 presents the estimates for independent normal distributions of the two random coefficients. The mean predictive power of deterministic intentions is similar to that estimated in the fixed coefficients model of Table 4a. and the associated standard deviation is small in size and not significantly different from zero. Hence, the model does not indicate substantial heterogeneity in the predictive power of deterministic intentions. The positive and significant mean coefficient on the interaction *intention* $\times$ *prob. questions* shows that on average, probabilistic intentions outperform deterministic ones. Moreover, the estimated standard deviation is significant as well and almost equally large as the mean, implying that for about 84% ($\Phi(0.0266/0.0266)$) of all respondents, the probability questions indeed provide additional power for predicting whether someone voted for a specific party. In the next subsection, we will analyse how this heterogeneity in the additional predictive power of subjective probabilities relates to characteristics of the individual. Probabilistic intentions have no additional predictive power for the "no vote" option, the placebo: the coefficient on the interaction *intention* $\times$ *prob. quest.* $\times$ *no vote* is negative and significant, cancelling the difference found for the other alternatives.

In order to facilitate interpretation of the magnitudes of the coefficients in Table 4, we report average marginal effects in Table 5. These are calculated as the average increase in the probability of voting for a given party that results from a 100pp increase in the intention to vote for that party. For each party we compare the situation in which the individual assigns a probability of zero to vote for this party and 1/8 to each other option with that in which (s)he assigns 100% to this party and zero to the other options. Table 5a uses the MNL estimates of Table 4a, allowing the predictive power of deterministic intentions and probabilities to vary

Table 5: Sample average marginal effect of 100pp increase in intent on probability of voting consistently

	Model 5a				Model 5b			
100pp increase in intent to vote...	Deter.	Prob.	Diff. (pp)[a]	Diff. (%)[b]	Deter.	Prob.	Diff. (pp)[a]	Diff. (%)[b]
... VVD (liberal)	0.74	0.87	13	18	0.66	0.79	13	20
... other party	0.41	0.71	30	73	0.60	0.80	20	33
... CDA (christian)	0.73	0.87	14	19	0.66	0.79	13	20
... D66 (prog. lib.)	0.62	0.85	23	37	0.64	0.80	16	25
... GL (green)	0.65	0.87	22	34	0.62	0.80	18	29
... PVV (populist)	0.58	0.75	17	29	0.50	0.77	27	54
... SP (socialist)	0.55	0.81	26	47	0.62	0.80	18	29
... PvdA (labour)	0.68	0.90	22	32	0.55	0.79	24	44
... no vote	0.45	0.41	-4	-9	0.45	0.39	-6	-13

Example: for the first alternative intentions change from $(0, 1/8, 1/8, ..., 1/8)$ to $(1, 0, 0, ..., 0)$.

Notes

[a] Difference probabilistic − deterministic.

[b] Percentage difference $^{(\text{prob.} - \text{deter.})}/_{\text{deter.}} \times 100$.

across the nine alternatives. The marginal effect of deterministic intentions is weakest for "other party" and strongest for the liberal VVD. Probabilities add 22-30pp for D66, GL, SP and "other party". These estimates corroborate the insights from the linear models reported above. The average marginal effects according to the mixed logit model in Table 5b are qualitatively similar and lead to the same overall conclusion, though the magnitudes of the differences between the two sub-samples are sometimes rather different.

4 Heterogeneity in the predictive power of probabilities

The mixed logit with normal mixing distributions presented in Table 4b can be used to back out estimates of the two individual specific parameters for each respondent. These individual specific estimates are the posterior means of the random coefficients, conditional on the individual's reported intentions and actual voting outcome. We are especially interested in the individual-specific parameter on the interaction *intention* × *prob. quest.* for the sample that received probabilistic questions, since this parameter provides a measure of the predictive

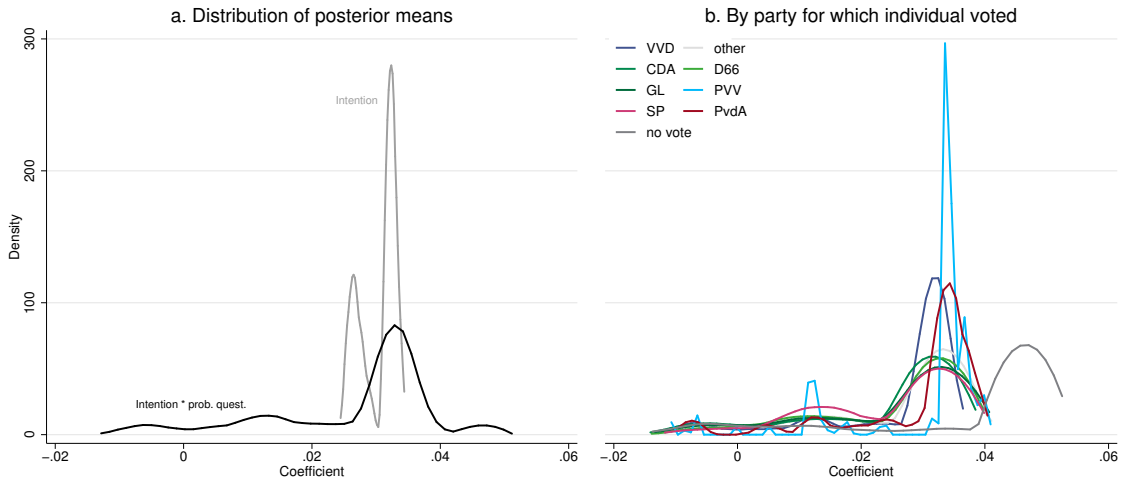

Figure 7: Kernel densities of random coefficients in mixed logit model 5b

power of subjective probabilities at the level of the individual.

The posterior means are proxies of the individual-specific parameters. There are two reasons why they are not identical to them. Firstly, the posterior means are calculated from the estimates reported in Table 4b and the estimation uncertainty of the mixed logit carries through in subsequent analysis. The online appendix analyses this source of estimation uncertainty in the individual-level parameters and explains how it can be accounted for. Estimation uncertainty of the mixed logit would disappear if the number of respondents tends to infinity. However, this still leaves the second issue: we only observe a single decision (the actual vote in the election) for each respondent. For any given individual the estimated posterior mean would be a consistent estimate of their parameter only if the number of observed choices would tend to infinity. In the analysis below, we just use the proxies (the posterior means) at face value and do not try to analyse their deviations from the individual specific parameters.

Figure 7 plots the densities of the posterior means for all respondents. Panel a. shows the distribution of the posterior means of the main effect of intentions and of the interaction of intentions and the probabilistic treatment dummy. We limit the sample to the relevant sub-

sample in both cases: respondents who received the deterministic questions for the main effect and respondents who received the probabilistic treatment for the interaction. As was evident from the estimates in Table 4b, there is little heterogeneity in the effect of deterministic intentions but substantial variation in the interaction term. The added value of probabilities thus varies across respondents: the density peaks for a coefficient just under 0.04 and it has a heavy left tail. The variation in coefficients indicates that the combination of a single set of intentions and a single vote already provides substantial information on the individual specific coefficients beyond the marginal information in the mixing distribution.

Panel b. of Figure 7 displays densities for the interaction term separately by the political party individuals voted for. It shows that the predominant feature of the overall density, the heavy left tail, is evident for the constituents of each party. The online appendix analyses estimation error in the posterior means. The results reported there indicate that sampling error in coefficients that results from using estimates of the mixed logit is substantial relative to the cross-sectional variation in the point estimates of posterior means. All estimates (and standard errors) reported below take this uncertainty into account.

Though the general shape of the distribution is similar for voters of all parties, the locations differ. For instance, Figure 7b shows that the distribution for respondents who did not vote lies slightly to the right of the other ones. Importantly, such variation in the posterior means of voters who choose different alternatives reflects features of the amount of information carried by the actual vote and the reported intentions in addition to variation in the extent to which intentions are consistent with actual decisions. For a given set of stated intentions, such as a 100% probability on the party that the individual actually voted for, there is significant variation in posterior means across parties. Therefore, we control for the party someone voted for in some specifications of the model discussed below, where the posterior means are analysed in relation to probability numeracy.

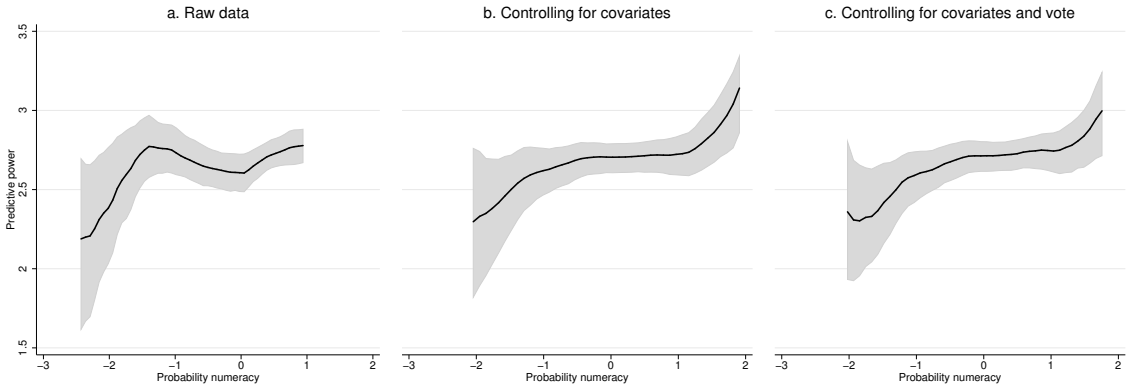

Figure 8: Kernel regressions of the predictive power of probabilistic intentions on probability numeracy (shaded areas are 95% confidence bands)

4.1 Probability numeracy and the predictive power of subjective probabilities

We construct a measure of probability numeracy from the 9-item scale that was administered as part of the 2008 *disease prevention* survey in the LISS panel. Appendix D contains a list of these items as well as the estimates of the item-response model used to aggregate them into a single measure for each individual. Unfortunately, we can only construct a numeracy score for half of our sample due to panel refreshments between 2008 and 2017.

Appendix E displays estimates of linear models that relate probability numeracy to background characteristics and actual voting behaviour. Probability numeracy varies significantly with voting behaviour: respondents who abstain from voting have the lowest average numeracy, followed by the socialists (SP) and populists (PVV). The other constituencies all have higher numeracy. Though education also clearly matters, with the higher educated displaying better numeracy, significant and substantial differences across parties remain if education is controlled for (see the second column of the table).

Figure 8 shows kernel regressions of the posterior means that measure the predictive power of subjective probabilities, the coefficients on *intention* $\times$ *prob. questions* multiplied

by 100, on probability numeracy. We find that the two are positively related and that this association becomes more pronounced when we control for demographics and for the party that an individual voted for.[5] The association is non-linear: at low levels of numeracy increases are associated with a tighter link between choice expectations and actual behaviour while the relationship flattens out for middle levels of numeracy and picks up again for those at the top end of the numeracy distribution.

Table 6 contains estimates of linear regression models of individual-level posterior means on numeracy, education and other controls (not reported). Numeracy enters the model linearly and is not transformed by taking the log, because it takes both positive and negative values. The estimates provide strong evidence that probability numeracy correlates positively with the extent to which probabilities predict voting. The association between numeracy and the predictive power of probabilities becomes stronger when controlling for education and the party one voted for. The strong link between numeracy and the predictive power of probabilities is especially striking in light of the eight year period between the elicitation of numeracy and the collection of voting data. Several robustness checks corroborate this significantly positive correlation between the predictive power of probabilities and probability numeracy, using both normal and log-normal mixing distributions and models with more random coefficients (on *intention* $\times$ *no vote* and *intention* $\times$ *prob. quest.* $\times$ *no vote*; results available upon request).

[5] Controlling for other covariates is achieved by first regressing probability numeracy on the other covariates and then performing the kernel regression of the posterior mean on the residual of the first regression rather than probability numeracy itself.

Table 6: Models of mixed logit interaction coefficients

	Dependent variable: mixed logit coef. × 100			
	(1)	(2)	(3)	(4)
Prob. numeracy	0.0753	0.125**	0.125**	0.154***
	(-0.0284; 0.197)	(0.000; 0.280)	(0.0198; 0.255)	(0.0308; 0.314)
Education (baseline: primary)				
Inter. secondary		-0.0364		-0.00202
		(-0.446; 0.355)		(-0.376; 0.377)
Higher secondary		-0.0559		0.0635
		(-0.540; 0.399)		(-0.387; 0.529)
Inter. vocational		-0.0473		0.0315
		(-0.466; 0.365)		(-0.349; 0.441)
Higher vocational		0.0188		0.102
		(-0.386; 0.429)		(-0.283; 0.519)
University		0.208		0.275
		(-0.260; 0.716)		(-0.167; 0.793)
Actual vote (baseline: no vote)				
VVD			-1.295***	-1.279***
			(-2.139; -0.579)	(-2.150; -0.521)
Other party			-1.046***	-1.151***
			(-1.796; -0.387)	(-2.017; -0.427)
CDA			-1.435***	-1.484***
			(-2.307; -0.677)	(-2.447; -0.664)
D66			-1.281***	-1.309***
			(-2.124; -0.548)	(-2.245; -0.530)
GL			-1.303***	-1.405***
			(-2.162; -0.551)	(-2.359; -0.596)
SP			-1.550***	-1.581***
			(-2.491; -0.733)	(-2.622; -0.634)
PVV			-0.837***	-0.882***
			(-1.580; -0.204)	(-1.680; -0.207)
PvdA			-0.968***	-0.853
			(-1.738; -0.317)	(-1.934; 3.660)
Controls	No	Yes[a]	No	Yes[a]
Observations	1002	1002	1002	1002
R-squared	0.0036	0.062	0.086	0.14

95% confidence intervals in parentheses.

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Confidence intervals and p-values take into account estimation error in the dependent variable (see online appendix).

Notes

[a] Specification controls for gender, age, net HH income, HH type (single; partner no children; partner with children; single with children; other), homeownership, urbanization, ethnicity (Dutch, 1st gen Western, 1st gen non-Western, 2nd gen Western, 2nd gen non-Western) and a full set of dummies for the party the respondent voted for.

5 Completion time and survey evaluation

In order to judge whether the additional predictive power of subjective probability questions makes incorporating them in the survey worthwhile, it seems useful to also consider the costs in terms of higher respondent burden. Each data set collected in the LISS panel includes variables that measure the time a respondent spent answering the questions, which can be seen as an approximation of the effort panel members put into their answers. In Table 7, we compare the time taken to complete the survey across the samples that received probabilistic and deterministic questions. Panel a. presents some percentiles (which are less sensitive to outliers than means and standard deviations), showing that a large majority of the respondents spend between 10 and 45 minutes on completing the survey, with a median slightly over 15 minutes. Interestingly, the percentiles are higher for the probabilistic than for the deterministic sample. Using quantile regressions, panel b. of Table 7 confirms that these differences are statistically significant and do not change much if we control for a wide range of demographics (as expected due to randomly assigned treatment). The probabilistic questions caused respondents to take 1 minute longer at the first quartile, 2 minutes at the median and 3 minutes at the third quartile. Apparently, respondents tend to put in more effort to report probabilities than they do to select a single party.

At the end of each survey, LISS routinely asks some diagnostic questions about the perceived difficulty and the extent to which respondents enjoyed filling out the questionnaire. We compared the answers across treatments, but did not find any significant differences between the two treatments; see Appendix B for details.[6] We can therefore conclude that even though the subjective probabilities required some additional effort, the respondents did not find the probabilistic survey substantially more difficult, less interesting, or less enjoyable than the deterministic one.

[6]This result might be due to the order of the questions; voting intentions (the only difference between the treatments) were located 55th and 56th among 170 items.

Table 7: Descriptive statistics and regression models of time to complete survey

a. Descriptive statistics of time (min.) to complete survey

		Percentiles				
	N	p10	p25	p50	p75	p90
Deterministic	1933	9.7	12.2	16.2	22.8	38.2
Probabilistic	2037	10.3	13.3	17.6	25.1	46.4

b. Quantile regressions of time (min.) to complete survey[a]

	Quantile regressions				
	p10	p25	p50	p75	p90
Prob. questions	0.72***	1.05***	1.81***	2.70***	7.31*
	(0.214)	(0.218)	(0.271)	(0.495)	(4.243)
Controls	Yes	Yes	Yes	Yes	Yes
N	3725	3725	3725	3725	3725

Standard errors in parentheses; clustered at household level (2867 clusters).
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Notes

[a] Specification controls for gender, age, education, net HH income, HH type (single; partner no children; partner with children; single with children; other), homeownership, urbanization and ethnicity (Dutch, 1st gen Western, 1st gen non-Western, 2nd gen Western, 2nd gen non-Western).

6 Conclusion

This paper looks at the predictive power of subjective probability questions in surveys that elicit how the same survey respondents intend to behave in the future. In particular, we use the context of parliamentary elections in the Netherlands to relate individuals' intentions to vote for different parties, elicited three months before the election, to how they actually voted. We exploit experimental variation in the question format used to measure intentions and compare deterministic items, in which respondents choose a single party as their best prediction, with probabilistic questions that allow individuals to express uncertainty and doubt. Such a probabilistic approach to polling has been applied to U.S. presidential elections since 2008, but we are the first to compare the predictive power of probabilistic and deterministic intentions using a large split-sample design. Moreover, while U.S. elections are contests between two parties, the Dutch elections ask voters to choose between many more options. Our

outcome variable distinguishes between nine possibilities: the seven major parties, the option to vote for any other party or to vote "blank", and the option not to vote at all. The multi-party nature of the Dutch political system creates scope for using probabilities to express undecidedness and indeed, 70% of the respondents assign positive probabilities to more than one party. On average, they assign positive probabilities to 2.6 parties. Such information is not contained in deterministic voting intentions, and our main research question is whether (and if so, for which respondents) it has added value for predicting individual behaviour.

Probabilistic questions may require more effort from respondents: at the median those who answered them took almost two minutes longer to complete the survey than respondents who were asked to choose a single party did. Nonetheless, both formats yield similar rates of item non-response and there are no significant differences between the two treatments in the evaluations of the difficulty or attractiveness of the survey. Comparing average intentions and actual votes at the aggregate level, we find that probabilities are closer to realized behaviour for 7 out of 9 options and the difference between question formats is significant for 4 of these.

Our main finding is that at the level of the individual, subjective probabilities are substantially better predictors of actual voting than deterministic intentions. This follows from non-parametric regressions as well as linear and non-linear regression models. In linear models, an increase from 0 to 100% in the deterministic intention to vote for a given party is associated with a 70pp increase in the likelihood of voting for that party. Probabilistic questions add another 20pp. Similarly, in multinomial logit and mixed logit models, the estimated average marginal effect of a probabilistic prediction is between 18 and 47 percent higher than that of a deterministic question for the same party. These benefits of probabilities over deterministic answers are largely due to the additional information contained in the probabilities: if probabilities are first discretized to resemble the deterministic answers (the party with the largest probability), their predictive performance is very similar to (and not significantly different from) that of the deterministic questions.

We use the estimates of a mixed logit model to approximate the predictive power of probabilistic intentions for each individual in the sub-sample that received those questions. Using normal mixing distributions we find that subjective probabilities provide additional predictive power for actual behaviour for a large majority (84%) of the respondents. There is a significantly positive association between the extent to which probabilities add to the predictive power and probability numeracy as measured using a 9-item scale. This must be a very persistent relation, since the measure of probability numeracy is constructed based on data from 2008 while the poll data was collected more than eight years later. The result is in line with earlier studies demonstrating heterogeneity in the ability to work with probabilities and their predictive value. On the other hand, we find that the predictive power of deterministic intentions hardly varies across the sample.

Comparing benefits and costs, our main overall conclusion is that subjective probabilities should be preferred over deterministic questions when one aims to predict behaviour at the level of the individual. They work well when predicting choices from a discrete menu of options. Future research can investigate whether the benefits observed for the application of voting extend to other domains.

References

Armantier, O., Bruine de Bruin, W., Topa, G., Van der Klaauw, W., and Zafar, B. (2015). Inflation expectations and behavior: Do survey respondents act on their beliefs? *International Economic Review*, 56(2):505–536.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Benjamini, Y. and Liu, W. (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference*, 82(1):163–170.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188.

Binswanger, J. and Salm, M. (2017). Does everyone use probabilities? The role of cognitive skills. *European Economic Review*, 98:73–85.

Bruine de Bruin, W. and Carman, K. G. (2012). Measuring risk perceptions: what does the excessive use of 50% mean? *Medical Decision Making*, 32(2):232–236.

Bruine de Bruin, W., Fischhoff, B., Millstein, S. G., and Halpern-Felsher, B. L. (2000). Verbal and numerical expressions of probability: "It's a fifty-fifty chance". *Organizational Behavior and Human Decision Processes*, 81(1):115–131.

Carman, K. G. and Kooreman, P. (2014). Probability perceptions and preventive health care. *Journal of Risk and Uncertainty*, 49(1):43–71.

De Vos, K. (2010). Representativeness of the liss-panel 2008, 2009, 2010. Mimeo CentERdata.

Delavande, A. and Manski, C. F. (2010). Probabilistic polling and voting in the 2008 presidential election evidence from the american life panel. *Public Opinion Quarterly*, 74(3):433–459.

Gutsche, T. L., Kapteyn, A., Meijer, E., and Weerman, B. (2014). The rand continuous 2012 presidential election poll. *Public Opinion Quarterly*, 78(S1):233–254.

Holland, B. S. and Copenhaver, M. D. (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics*, 43(2):417–423.

Hurd, M. (2009). Subjective probabilities in household surveys. *Annual Economic Review*, 1:543–562.

Kleinjans, C. and Van Soest, A. (2014). Non-response and focal point answers to subjective probability questions. *Journal of Applied Econometrics*, 29(4):567–585.

Maas, K., Steenbergen, M., and Saris, W. (1990). Vote probabilities. *Electoral Studies*, 9(2):91–107.

Manski, C. and Molinari, F. (2010). Rounding probabilistic expectations in surveys. *Journal of Business Economics and Statistics*, 28(2):219–231.

Manski, C. F. (2004). Measuring expectations. *Econometrica*, 75(5):1329–1376.

Marks, G., Hooghe, L., Nelson, M., and Edwards, E. (2006). Party competition and European integration in the East and West. *Comparative Political Studies*, 39(2):155–175.

Meier, K. J. (1980). Rationality and voting: A downsian analysis of the 1972 election. *Western Political Quarterly*, 33(1):38–49.

Meier, K. J. and Campbell, J. E. (1979). Issue voting: an empirical examination of individually necessary and jointly sufficient conditions. *American Politics Quarterly*, 7(1):21–50.

Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Annals of Statistics*, 30(2):239–257.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754.

Train, K. (2003). *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge.

Van der Laan, J. (2009). Representativity of the LISS panel. Statistics Netherlands discussion paper 09041.

Van Kersbergen, K. and Krouwel, A. (2008). A double-edged sword! The Dutch centre-right and the 'foreigners issue'. *Journal of European Public Policy*, 15(3):398–414.

Appendix A: balance tests

Table A1: Balance tests: descriptive statistics of covariates by question type

	Overall						
	N	Mean	SD	Deterministic	Probabilistic	Difference	(SE)
Female	3978	0.51	0.50	0.51	0.50	-0.01	(0.016)
HH members	3978	2.4	1.3	2.4	2.5	0.07*	(0.039)
Partner	3978	0.70	0.46	0.69	0.72	0.03*	(0.015)
Age	3978	54	17	54	54	-0.1	(0.54)
Net HH income	3882	3108	3574	3148	3070	-78	(116.0)
Homeowner	3938	0.73	0.44	0.74	0.73	-0.01	(0.014)
Prob. numeracy[a]	2033	0.00	0.84	-0.04	0.04	0.07*	(0.038)
Education							
Primary	3974	0.06	0.24	0.06	0.06	-0.01	(0.007)
Inter. secondary	3974	0.22	0.41	0.23	0.21	-0.02*	(0.013)
Higher secondary	3974	0.11	0.31	0.10	0.12	0.02**	(0.010)
Inter. vocational	3974	0.24	0.43	0.25	0.23	-0.02	(0.014)
Higher vocational	3974	0.26	0.44	0.25	0.26	0.02	(0.014)
University	3974	0.12	0.32	0.11	0.12	0.01	(0.010)
Ethnicity							
Dutch	3872	0.87	0.34	0.86	0.88	0.02**	(0.011)
First gen. Western	3872	0.03	0.16	0.03	0.03	-0.001	(0.005)
First gen. non-Western	3872	0.03	0.18	0.04	0.03	-0.002	(0.006)
Sec. gen. Western	3872	0.05	0.22	0.05	0.05	-0.007	(0.007)
Sec. gen. non-Western	3872	0.02	0.14	0.03	0.01	-0.01***	(0.004)
Urbanisation							
Extremely	3974	0.15	0.35	0.15	0.14	-0.006	(0.011)
Very	3974	0.26	0.44	0.26	0.26	0.004	(0.014)
Moderately	3974	0.22	0.42	0.23	0.21	-0.02	(0.013)
Slightly	3974	0.22	0.41	0.21	0.22	0.003	(0.013)
Rural	3974	0.16	0.36	0.14	0.17	0.02*	(0.012)

Reported statistical significance is not corrected for multiple comparisons. No single null-hypothesis is rejected if we correct the initial p-cutoff of 0.05 for multiple comparisons using the methods proposed by Holland and Copenhaver (1987); by Benjamini and Liu (1999) and Sarkar (2002); or by Simes (1986), Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001). Only a single null-hypothesis is rejected using these methods for an initial p-cutoff of 0.10, corresponding to second generation non-Western migrants. Balance checks correcting for multiple comparisons are available on request.

Standard errors in parentheses; clustered at household level.

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Notes

[a] Computed using one-parameter logistic item response model; estimates reported in Appendix D.

Appendix B: question difficulty

Table B1: Question difficulty

a. Descriptive statistics of survey evaluation					
Questions were...	... difficult	... clear	... thought-provoking	... interesting	... enjoyable
1. Certainly not	0.47	0.01	0.10	0.03	0.02
2.	0.19	0.02	0.11	0.04	0.04
3.	0.17	0.14	0.36	0.29	0.32
4.	0.12	0.35	0.27	0.36	0.32
5. Certainly yes	0.05	0.48	0.17	0.28	0.30
N	3970	3970	3970	3970	3970
b. Ordered logit models of survey evaluation[a]					
Questions were...	... difficult	... clear	... thought-provoking	... interesting	... enjoyable
Prob. questions	0.046	-0.072	-0.075	-0.108*	-0.072
	(0.062)	(0.063)	(0.059)	(0.061)	(0.0613)
Controls	Yes	Yes	Yes	Yes	Yes
N	3725	3725	3725	3725	3725
LLH	-5044.67	-4138.71	-5499.29	-4818.51	-4740.34

Standard errors in parentheses; clustered at household level (2867 clusters).
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Notes

[a] Specification controls for gender, age, education, net HH income, HH type (single; partner no children; partner with children; single with children; other), homeownership, urbanization and ethnicity (Dutch, 1st gen Western, 1st gen non-Western, 2nd gen Western, 2nd gen non-Western).

Appendix C: linear models of actual vote with covariates

Table C1: Linear probability models of actual vote (0% or 100%) with covariates

a. Baseline (controls included in the models)

		Parties							
	no vote	VVD	PVV	CDA	D66	GL	SP	PvdA	other
Intention (%)	0.418***	0.785***	0.565***	0.760***	0.652***	0.664***	0.558***	0.692***	0.444***
	(0.0326)	(0.0251)	(0.0331)	(0.0337)	(0.0387)	(0.0420)	(0.0459)	(0.0413)	(0.0303)
Intention × prob. questions	-0.0706	0.166***	0.119**	0.159***	0.207***	0.222***	0.232***	0.157***	0.275***
	(0.0466)	(0.0401)	(0.0480)	(0.0490)	(0.0582)	(0.0630)	(0.0663)	(0.0592)	(0.0442)
Prob. questions	-0.0183	-2.282**	-0.850*	-3.388***	-1.877**	-2.043***	-2.034***	-1.982***	-2.097**
	(0.574)	(0.892)	(0.483)	(0.846)	(0.790)	(0.681)	(0.734)	(0.533)	(0.827)
N	3732	3732	3732	3732	3732	3732	3732	3732	3732
R-squared	0.25	0.48	0.44	0.39	0.36	0.36	0.29	0.45	0.28

b. Probabilistic intentions replaced by mode (controls included in the models)

		Parties							
	no vote	VVD	PVV	CDA	D66	GL	SP	PvdA	other
Intention (%)	0.418***	0.781***	0.564***	0.758***	0.648***	0.660***	0.557***	0.693***	0.444***
	(0.0326)	(0.0251)	(0.0331)	(0.0337)	(0.0386)	(0.0420)	(0.0459)	(0.0414)	(0.0303)
Intention × prob. questions	-0.0706	-0.0277	0.0394	-0.0111	-0.0134	0.0572	0.0637	-0.0104	0.154***
	(0.0466)	(0.0362)	(0.0463)	(0.0449)	(0.0516)	(0.0562)	(0.0621)	(0.0554)	(0.0430)
Prob. questions	-0.0183	-0.517	-0.165	-1.941**	0.544	-0.590	-0.675	-0.593	0.576
	(0.574)	(0.920)	(0.470)	(0.862)	(0.787)	(0.689)	(0.736)	(0.528)	(0.831)
N	3732	3732	3732	3732	3732	3732	3732	3732	3732
R-squared	0.25	0.46	0.44	0.38	0.35	0.37	0.28	0.45	0.27

Specification controls for gender, age, education, net HH income, HH type (single; partner no children; partner with children; single with children; other), homeownership, urbanization and ethnicity (Dutch, 1st gen Western, 1st gen non-Western, 2nd gen Western, 2nd gen non-Western).
Systems of linear probability models are estimated as SUR models (allowing for dependence between error terms of different equations). Robust standard errors in parentheses, clustered at household level (2871 clusters).
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Appendix D: estimates of item response model for probability numeracy

Table D1: Estimates of item response model used to predict probability numeracy

a. Items ranked by increasing difficulty

Item	Question
q1	If the chance of getting a disease is 10%, how many people out of 100 would be expected to get the disease?
q2	If the chance of getting a disease is 10%, how many people out of 1000 would be expected to get the disease?
q3	Which of the following represents the biggest risk of getting a disease? 1%; 10%; 5%
q4	Which of the following numbers represents the biggest risk of getting a disease? 1 in 100; 1 in 1000; 1 in 10
q5	In the BIG BUCKS LOTTERY, the chances of winning a $10.00 prize are 1%. What is your best guess about how many people would win a $10.00 prize if 1,000 people each buy a single ticket from BIG BUCKS?
q6	If the chance of getting a disease is 20 out of 100, this would be the same as having a ... % chance of getting the disease.
q7	Imagine that we roll a fair, six-sided die 1000 times. Out of 1000 rolls, how many times do you think the die would come up even (2, 4, or 6)?
q8	The chance of getting a viral infection is .0005. Out of 10,000 people, about how many of them are expected to get infected?
q9	In the ACME PUBLISHING SWEEPSTAKES, the chance of winning a car is 1 in 1000. What percent of tickets of ACME PUBLISHING SWEEPSTAKES win a car? ... % of the tickets

b. Estimates of item response model

		Item-specific parameters								
		Easy								Difficult
		q1	q2	q3	q4	q5	q6	q7	q8	q9
Discrimination	1.742***									
	(0.0501)									
Difficulty		-2.079***	-1.922***	-1.371***	-1.142***	-1.066***	-0.941***	-0.571***	-0.460***	-0.214***
		(0.0788)	(0.0723)	(0.0535)	(0.0501)	(0.0485)	(0.0481)	(0.0427)	(0.0418)	(0.0409)
Individuals		2045								
Log-likelihood		-8111.44								

Robust standard errors in parentheses, clustered at household level (1589 clusters).
***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

Appendix E: linear models of probability numeracy

Table E1: OLS regression models of probability numeracy

	Probability numeracy	
Actual vote (baseline: no vote)		
VVD	0.782***	0.356***
	(0.122)	(0.122)
Other party	0.642***	0.380***
	(0.124)	(0.117)
CDA	0.638***	0.320***
	(0.127)	(0.123)
D66	0.808***	0.353***
	(0.130)	(0.129)
GL	0.854***	0.483***
	(0.138)	(0.133)
SP	0.266*	0.165
	(0.142)	(0.132)
PVV	0.353**	0.240*
	(0.138)	(0.124)
PvdA	0.783***	0.542***
	(0.137)	(0.133)
Education (baseline: primary)		
Inter. secondary		0.316**
		(0.126)
Higher secondary		0.684***
		(0.142)
Inter. vocational		0.485***
		(0.130)
Higher vocational		0.750***
		(0.132)
University		0.929***
		(0.138)
Controls	No	Yes[a]
Observations	1040	1002
R-squared	0.082	0.317

Cluster-robust standard errors in parentheses (921 and 885 HHs)
***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

Notes

[a] Specification controls for gender, age, net HH income, HH type (single; partner no children; partner with children; single with children; other), homeownership, urbanization and ethnicity (Dutch, 1st gen Western, 1st gen non-Western, 2nd gen Western, 2nd gen non-Western).

Online appendix: simulation procedure for random parameters

We use the estimates of the mixed logit model in Table 4b to compute posterior means conditional on reported choice expectations (intentions) and actual decisions. Therefore, we carry estimation error from the mixed logit through in the linear models of posterior means. We do this by drawing $S = 5000$ vectors of mixed logit coefficients from their asymptotic distribution and computing posterior means for each of them. Denote the parameters of the mixed logit model by $\boldsymbol{\theta}$ and the individual-specific parameter for respondent n by β_n. The simulations proceed as follows (see Train, 2003, for details):

1. **Outer loop.** For $s = 1, 2, ..., S$: draw $\boldsymbol{\theta}^s$ from its estimated sampling distribution $N\left(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{V}}\right)$.

2. **Inner loop.** For each s: simulate $\check{\beta}_n|_{\boldsymbol{\theta}=\boldsymbol{\theta}^s}$ as

$$\check{\beta}_n|_{\boldsymbol{\theta}=\boldsymbol{\theta}^s} = \sum_{r=1}^{R} w^r \beta^r|_{\boldsymbol{\theta}=\boldsymbol{\theta}^s}$$

   where $w^r = \frac{\Pr(y_n|\mathbf{x}_n,\beta^r)}{\sum_{r=1}^{R}\Pr(y_n|\mathbf{x}_n,\beta^r)}$ and $\beta^r|_{\boldsymbol{\theta}=\boldsymbol{\theta}^s}$ is drawn from the mixing distribution at the parameter vector $\boldsymbol{\theta}^s$. E.g. for the normal mixing distribution we draw $\beta^r|_{\boldsymbol{\theta}=\boldsymbol{\theta}^s}$ from $N(\mu^s, \sigma^s)$. We set R, the number of simulations in the inner loop, to 500.

Halton draws based on different primes are used to achieve variance reduction in both loops.

The values of $\check{\beta}_n|_{\boldsymbol{\theta}=\boldsymbol{\theta}^s}$ for the S parameter vectors $\boldsymbol{\theta}^s$ are draws from the sampling distribution of $\check{\beta}_n$ induced by the sampling distribution of $\widehat{\boldsymbol{\theta}}$. We summarize this estimation noise in Table OA1, which contains descriptive statistics of individual-level parameters evaluated at the mixed logit estimates, $\check{\beta}_n|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$, and of the mean and standard deviation of $\check{\beta}_n|_{\boldsymbol{\theta}=\boldsymbol{\theta}^s}$ across simulation draws s. The mean sampling variability of the coefficient on *intention* is large

Table OA1: Descriptives of individual-level parameters and their sampling distributions (normal mixing distribution)

		Simulated parameter[a]		Mean across draws[b]		SD across draws[c]	
	N	Mean	SD	Mean	SD	Mean	SD
Intention	1936	0.030	0.0022	0.030	0.0035	0.0032	0.00082
Intention × prob. quest.	2042	0.027	0.0093	0.027	0.013	0.0039	0.0012

[a] $\check{\beta}_n|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$
[b] $\bar{\beta}_n = \frac{1}{S}\sum_{s=1}^{S} \check{\beta}_n|_{\boldsymbol{\theta}=\boldsymbol{\theta}^s}$
[c] $\text{SD}\left(\check{\beta}_n\right) = \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}\left(\check{\beta}_n|_{\boldsymbol{\theta}=\boldsymbol{\theta}^s} - \bar{\beta}_n\right)^2}$

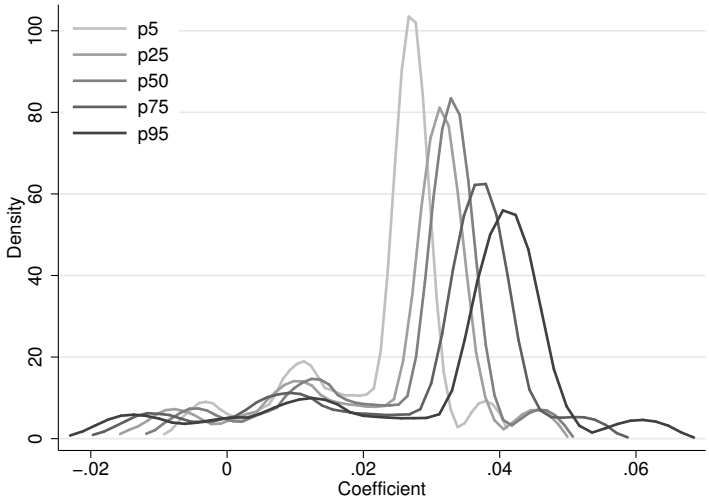

Figure OA1: Densities of random coefficient on *intention* × *prob. quest.* at various percentiles of the mean coefficient

relative to the variation in coefficients across respondents. This can be seen by comparing the SD of the mean coefficient across the sample (0.0035) with the mean across the sample of the SD across simulation draws (0.0032). The interaction with the indicator for probabilistic questions yields a sampling noise SD of 0.0039 and a cross-sample SD of 0.013. Though this is a much better signal to noise ratio, it remains potentially important to incorporate estimation error of the mixed logit parameters into subsequent inference.

Figure OA1 visualizes noise in the cross-sectional distribution of random coefficients on the interaction by plotting their density for five different parameter vectors $\boldsymbol{\theta}^s$. In order to

ensure that these densities adequately reflect sampling variation, we pick parameter vectors based on percentiles of the cross-sectional average coefficient. The lightest line labeled "p5", for instance, corresponds to the first stage parameter vector $\boldsymbol{\theta}^s$ that yields the 250th smallest sample average coefficient out of the 5000 draws. The cross-sectional distributions all have a similar shape with a heavy left tail. Their location shifts as we move through the distribution of the mean coefficient, the magnitude of the difference between the peaks of p5 and p95 is about 0.02.

The evidence presented above indicates that estimation error in posterior means is substantial relative to cross-sectional variation in their point estimates. However, lack of stability of the cross-sectional distribution of $\check{\beta}_n|_{\boldsymbol{\theta}=\boldsymbol{\theta}^s}$ does not imply that coefficients of linear models that explain those coefficients will themselves be unstable. Estimating such models requires one to take into account an additional source of variation. As long as we limit ourselves to functions of $\check{\beta}_n$, the only relevant uncertainty in our estimate $\check{\beta}_n$ is that which results from the sampling error in $\widehat{\boldsymbol{\theta}}$. There is an additional source of noise to account for when we use $\check{\beta}_n$ as dependent variable in a statistical model. Not only we do not observe our dependent variable perfectly, but we should also factor in the usual estimation uncertainty from observing a random sample of respondents. We do this by taking a different bootstrap sample for each draw $\check{\beta}_n|_{\boldsymbol{\theta}=\boldsymbol{\theta}^s}$ and calculating the parameters of the second stage model from that combination of $\check{\beta}_n|_{\boldsymbol{\theta}=\boldsymbol{\theta}^s}$ and the bootstrap sample. The bootstrap is clustered at the household level. For a given respondent n we capture estimation noise in the first stage by variation in $\check{\beta}_n|_{\boldsymbol{\theta}=\boldsymbol{\theta}^s}$ across $\boldsymbol{\theta}^s$. Different bootstrap samples contain different combinations of individuals to approximate sampling error in the second stage model. We denote the S estimators of the second stage parameters $\boldsymbol{\gamma}$ as $\widehat{\boldsymbol{\gamma}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^s}$.

Figure OA2 shows the density over bootstrap samples and $\boldsymbol{\theta}^s$ of the coefficient on probability numeracy for the linear models reported in Table 6 of the main text. The densities capture the combined variation of the first stage estimation of $\widehat{\boldsymbol{\theta}}$ and sampling error in the

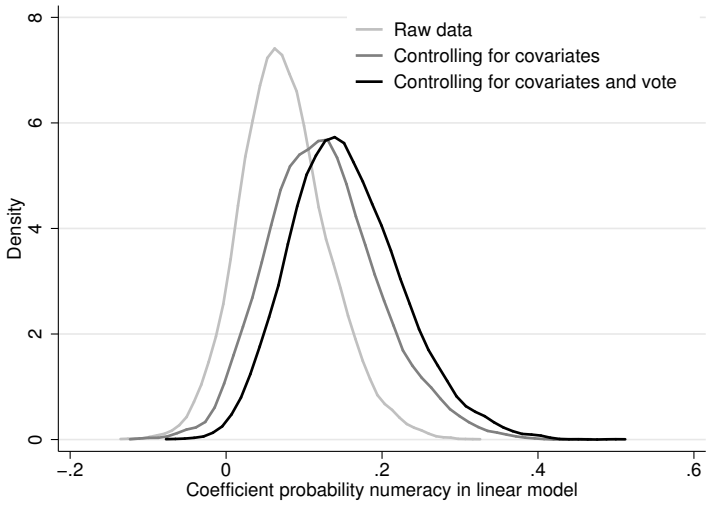

Figure OA2: Density of coefficients on probability numeracy in second stage linear model

linear model. Since all three sampling distributions are heavily skewed, we do not use any normal approximations to compute p-values and confidence intervals. Instead, we calculate these quantities directly from the simulation draws and report 95% confidence intervals rather than standard errors in Table 6 of the main text.